Data Science Project Proposal

Project Title	**Black Friday Data Set: Predict top5 product items most likely to be purchased by a consumer**
Proposed by	2017320160 임혜수 2017320214 이지원 2017320233 김채령
Submitted date	2019/4/13

Project Description

A. Dataset - Black Friday Data

The dataset here is a sample of the transactions made in a retail store on a Black Friday. With 550,000 observations, it contains different kinds of variables either numerical or categorical. It contains missing values. The data source is given in a csv format having 538K rows and 12 columns. The following describes each column with its name, containing information and given data type.

User_ID	User, numeric
Product_ID	ID Product, numeric (3623 unique values)
Gender	gender of a customer, categorical (boolean)
Age	age of a customer, categorical (range)
Occupation	ID Occupation of each customer, numeric
City_Category	current staying in city, categorical
Stay_In_Current_City_Years	years stayed in the city for each customer, categorical (range)
Marital_Status	marital status, categorical
Product_Category_1	product category 1, numeric
Product_Category_2	product category 2, numeric
Product_Category_3	product category 3, numeric
Purchase	purchase amount in dollars, numeric

Note. Black Friday – A study of sales through consumer behaviours by Mehdi Dagdoug, retrieved from "https://www.kaggle.com/mehdidag/black-friday"

B. Problem - Predict top 5 most likely to be purchased products for customers.

Using the Black Friday Dataset with 538K instances and 12 attributes, we'd like to expect top 5 products to a specific customer for the coming Black Friday. Here, 'top 5 products' are the five products that are most likely to be purchased by the customer. We are planning to build collaborative filtering models for expecting product items. Furthermore, with the prediction result we aim to recommend users to put top 5 items into their shopping cart.

C. Expected Result

1) Functions that search for an expected list based on a specified user
    - Input : User_ID
    - Output : ranked list of items (Product_IDs) that the user most likely to buy

2) Expected Usage

Firstly, we can predict the product items that are likely to be purchased by the customers. With these predicted items, we can also recommend items to customers. Then, we can adopt the system we implemented to the online shopping mall of the retail shop to make customers put the recommended items into their shopping carts. Plus, it could be also used to predict sales amount afterwards, and the analysis of the data might be able to suggest some meaningful marketing strategies by using statistical methods. In addition, correlation analysis is under our contemplation. We are going to figure out the correlation between given features, and if there exists any meaningful outcome, it's also applicable to products display order. All of these are expected to bring out positive contribution to market sales on the next Black Friday.

Project Tasks

A. Tasks details & Outline

1. Data Analysis : In order to understand our data, we are going to analyze data by visualizing. The goal of this step is to figure out how data is distributed and understand the meaning of each attribute.

2. Data Preprocessing : We are going to clean the data by handling missing values, detecting outliers, getting rid of redundant attributes, and dealing with errors and

artifacts.

3. Correlation Analysis : Using clean data, we are going to analyze correlation between attributes. This step is aim to find out important relationships between features. We are planning to use xgboost model. (This is only plan so the model we use can be changed.)

4. Build Recommendation System : We are going to build recommendation system using item based collaborative filtering. Rough outline is as below.
    a. Import Modules
    b. Load Data
    c. Data preparation
    d. Data split
    e. Baseline Model
    f. Collaborative Filtering Model (item based)
        - Define similarity function
    g. Model Evaluation
    h. Model Selection
    i. Final Output
        - CSV output file
        - Customer recommendation function
    j. Summary

5. Summary : We are going to understand how data is organized by analyzing dataset at step 1, then clean the data at step 2. By computing correlation between features at step 3, we are going to analyze how features are related to each other. At step 4, we are going to build recommendation system using item based collaborative filtering so that we can predict top 5 products that user is likely to buy.

6. Discussion : Through the outcomes of our analysis, we can utilize the results to products display order. Plus, it's also possible to recommend product items to users and predicting sales amount seems to be a feasible scenario with the price information afterwards.

B. Role

임혜수 : Data Analysis. Data Interpretation.

이지원 : Documentation. Repository management.

김채령 : Model Building. Model Implementation.